

LA-UR-21-21280

Approved for public release; distribution is unlimited.

Title: Spectra Swarm Evaluation

Author(s): Thompson, James Russell

Intended for: Report

Issued: 2021-02-11

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Spectra Swarm Evaluation

October 2020

Jim Thompson
HPC Systems - Archive Administrator

Acknowledgments:

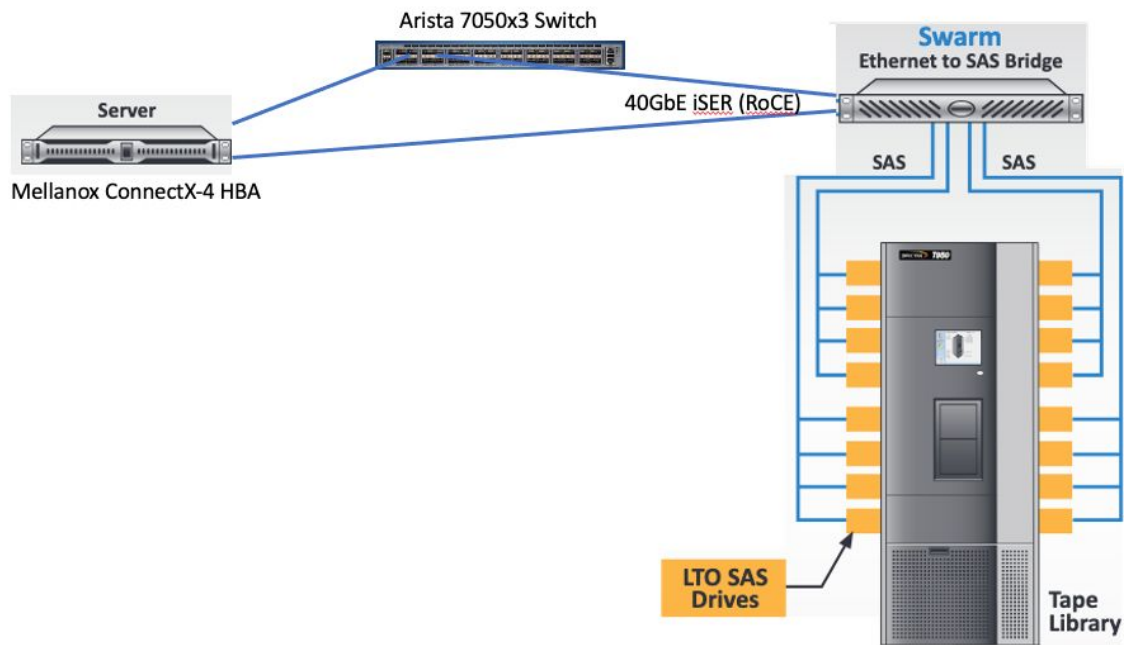
Brett Holander
Chase Harrison
Bob Darlington
Terri Bednar

Overview	3
Evaluation Environment / Equipment	4
Swarm	4
Host	4
Switch	5
Tape Library	6
Configuration Steps	7
iSCSI Target Management	7
Swarm Manual Target Mgmt	7
iSCSI Mapping	7
Redhat configuration commands	9
Performance Results	10
Single stream	10
Drive Average Results	11
Direct connect results	12
Switch connect results	13
Switch vs Direct	13
Max_sectors 1024 vs 2048	14
Spectrum Protect	15
Summary	16
Performance	16
Configuration	16
Troubleshooting	16
Redundancy	16
Cost benefit analysis	16

Overview

Spectra Swarm product is an ethernet to Serial-attached-SCSI(SAS) bridge that allows a host server to communicate with SAS tape drives over ethernet using RDMA over Converged Ethernet (RoCE).

Evaluation Environment / Equipment



Swarm

ATTO XstreamCORE ET 8200T Firmware: 4.02.002b Base Version: 3.05 SN: et8200t10XXXX

Host

PowerEdge R720 (40 logical cpus, 256 GB RAM)
2 x Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz (10 cores, 20 threads)

RHEL 7.8 - 3.10.0-1127.8.2.el7.x86_64

Mellanox Technologies MT27700 Family [ConnectX-4]

Part number: MCX414A-GCAT

fw_ver: 12.28.1002

Device Type: ConnectX4

Part Number: MCX414A-GCA_Ax

Description: ConnectX-4 EN network interface card; 50GbE dual-port QSFP28; PCIe3.0 x8; ROHS R6

PSID: MT_2610110035
PCI Device Name: /dev/mst/mt4115_pciconf0
Base GUID: b8599f03001afe8e
Base MAC: b8599f1afe8e
Versions: Current Available
 FW 12.28.1002 N/A
 PXE 3.6.0101 N/A
 UEFI 14.21.0016 N/A

mlnx-ofa_kernel-modules-5.1-OFED.5.1.0.6.6.1.kver.3.10.0_1127.8.2.el7.x86_64.x86_64

Switch

Arista DCS-7050CX3-32S-R Software image version: 4.23.4.2M

```
platform trident mmu queue profile RoCELosslessProfile
  ingress threshold 1/16
  egress unicast queue 3 threshold 8
!
port-channel load-balance trident fields mac src-mac dst-mac
port-channel load-balance trident fields ip source-ip destination-ip
source-port
!
interface Ethernet1/1
  description SWARM-HOST1 PORT A
  mtu 1514
  dcbx mode ieee
  flowcontrol send on
  flowcontrol receive on
  speed forced 40gfull
  switchport access vlan 100
!
interface Ethernet6/1
  description SWARM APPLIANCE
  mtu 1514
  dcbx mode ieee
  flowcontrol send on
  flowcontrol receive on
  speed forced 40gfull
  switchport access vlan 100
!
```

Tape Library

Library Name: tsm-t950a

BlueScale12.7.06.02

BlueScale12.7.06.02-20190719F

IBM Ultrium-TD8 Half Height Serial Attached SCSI

Drive FW: K4K1

DCM FW: 7.6.2

Configuration Steps

iSCSI Target Management

iSCSI TARGET MANAGEMENT

For the default target and manually-created targets, Access Control and CHAP are disabled.

Configure iSCSI Targets

default	Access Control	Device Maps	iSCSI CHAP	
21610090a5003287	Access Control	Device Maps	iSCSI CHAP	Delete disabled due to initiator login(s)
Discovery			iSCSI CHAP	

Add an iSCSI target :

SUBMIT

Swarm Manual Target Mgmt

Powered by
ATTO

HOME | BACK

STATUS

ETHERNET

TIME & DATE

REMOTE MANAGEMENT

ISCSI

MANUAL TARGET MGT

CONTROLLER

CERTIFICATE MANAGEMENT

FIRMWARE UPDATE

ADVANCED

RESTART

HELP

ACCESS CONTROL

FOR 21610090a5003287

Access Control : ☒ enabled ☐ disabled

List of Initiators

▼

▲

Allowed Initiators

iqn.1994-05.com.redhat:87465ac4ae86

iSCSI Mapping

iSCSI MAPPING

FOR

LUN 0 Controller	LUN 1 Default Target LUN : 1	LUN 2 Default Target LUN : 2
LUN 3 Default Target LUN : 3	LUN 4 Default Target LUN : 4	LUN 5 Default Target LUN : 5
LUN 6 Default Target LUN : 6	LUN 7 Default Target LUN : 7	LUN 8 Default Target LUN : 8
LUN 9 Default Target LUN : 9	LUN 10 Default Target LUN : 10	LUN 11 Default Target LUN : 11
LUN 12 Default Target LUN : 12	LUN 13 Default Target LUN : 13	LUN 14 Default Target LUN : 14
LUN 15 Default Target LUN : 15	LUN 16 Default Target LUN : 16	LUN 17 Default Target LUN : 17
LUN 18	LUN 19	LUN 20

Unmapped Devices

Default Target LUN : 18 OFFLINE

Redhat configuration commands

Check communication and iscsi target set up on swarm.

```
Host: iscsiadm -m discovery -t st -p 10.1.1.10
10.1.1.10:3260,1 iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:default
10.2.1.10:3260,1 iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:default
10.1.1.10:3260,1 iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:21610090a5003287
10.2.1.10:3260,1 iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:21610090a5003287
```

Logout and remove devices

```
iscsiadm --mode node --logoutall=all
```

Set transport to iser. Otherwise defaults to TCP.

```
Host: iscsiadm -m node -T iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:21610090a5003287 -o
update -n iface.transport_name -v iser
```

Discover devices sets device special files in /dev.

```
Host: iscsiadm -m node -T iqn.2016-10.com.atto:xcoreet:sn-et8200t10XXXX:0:21610090a5003287 -p
10.1.1.10 -l
```

Check from Swarm that initiators are logged in with the correct transport.

Swarm CLI: displayinitiators

```
; Port : Initiator : Target : Transport
;=====
1 DP1 : iqn.1994-05.com.redhat:87465ac4ae86 : 21610090a5003287 :
ISER
1 DP2 : iqn.1994-05.com.redhat:87465ac4ae86 : 21610090a5003287 :
ISER
```

How to check the iscsi initiatorname from the host.

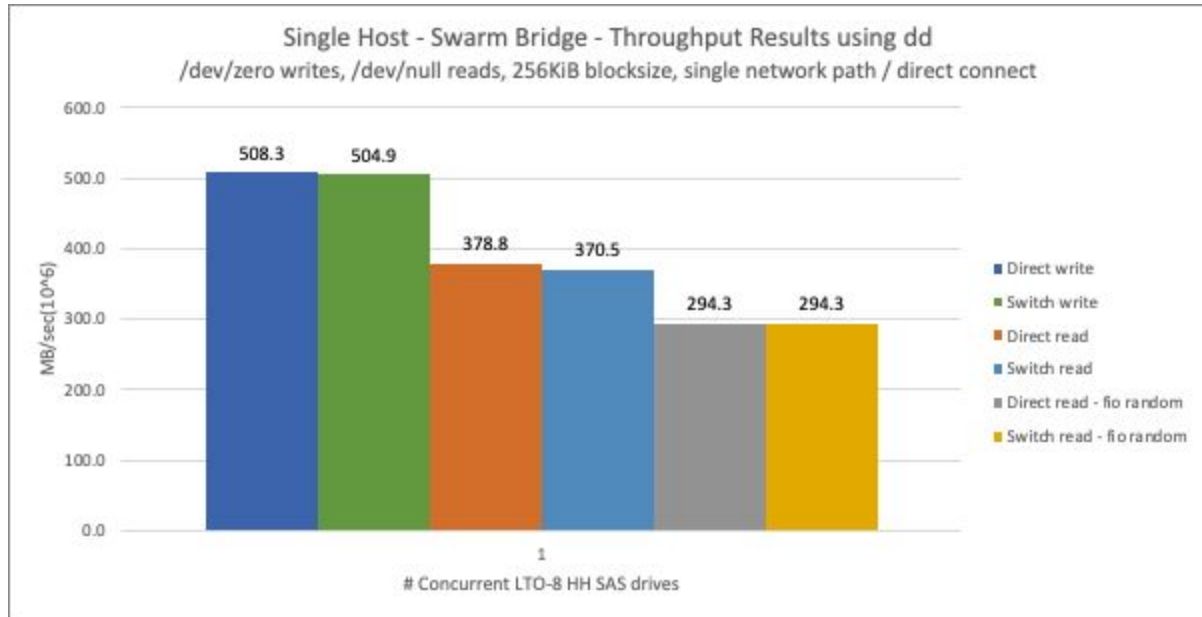
```
Host: cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1994-05.com.redhat:87465ac4ae86
```

How to set max_sectors for the ib_iser module.

```
Host: cat /etc/modprobe.d/ib_iser.conf
options ib_iser max_sectors=2048
```

Performance Results

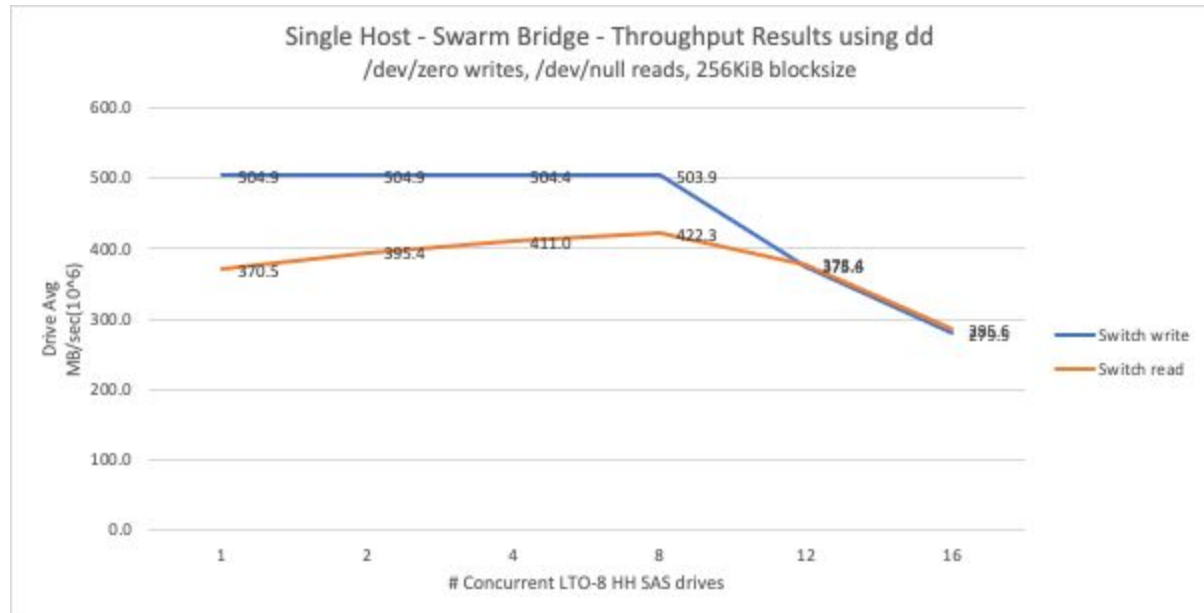
Single stream



No significant difference between switch and direct attachment. Fio random data points are from filling up the tape with random data using the fio tool, then reading with dd and sending to /dev/null. Likely that using fibre channel or SAS without the bridge would result in higher achievable compressed data rates.

Read vs write performance difference should be investigated. Tape drive would not be the 100+ MB/sec difference between writes and reads.

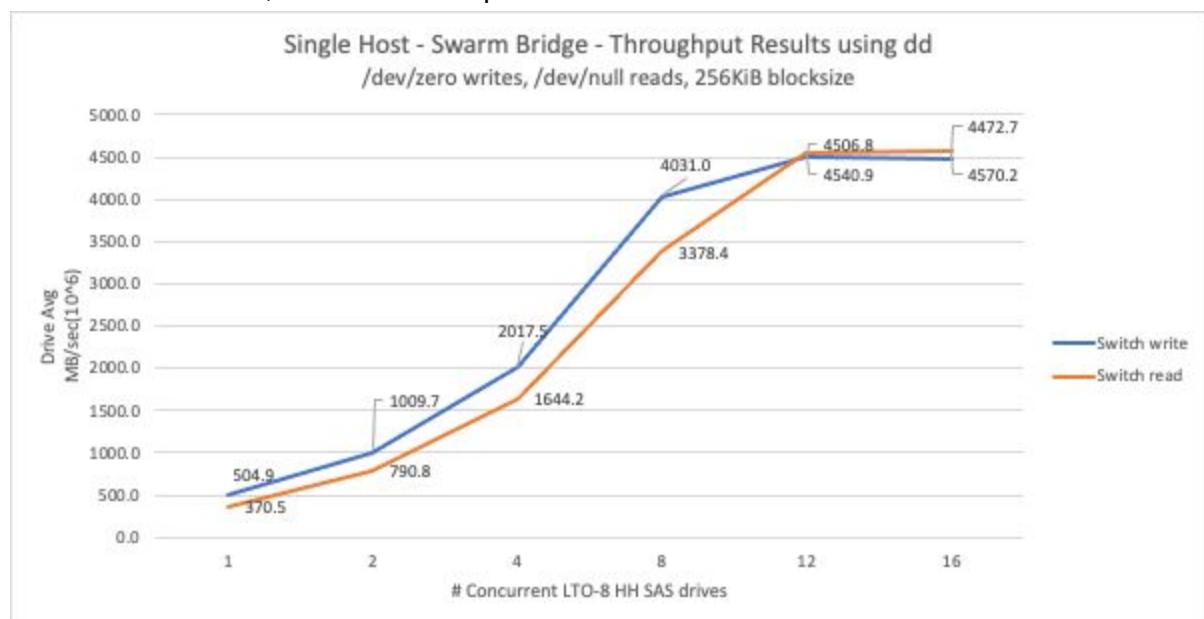
Drive Average Results



This chart shows the drive average MB/sec as the # of concurrent drives is scaled. Writes follow expected/desired behaviour. As the # of concurrent drives increases, initially the performance scaled linearly. At some point a system bottleneck is encountered and the average drops off.

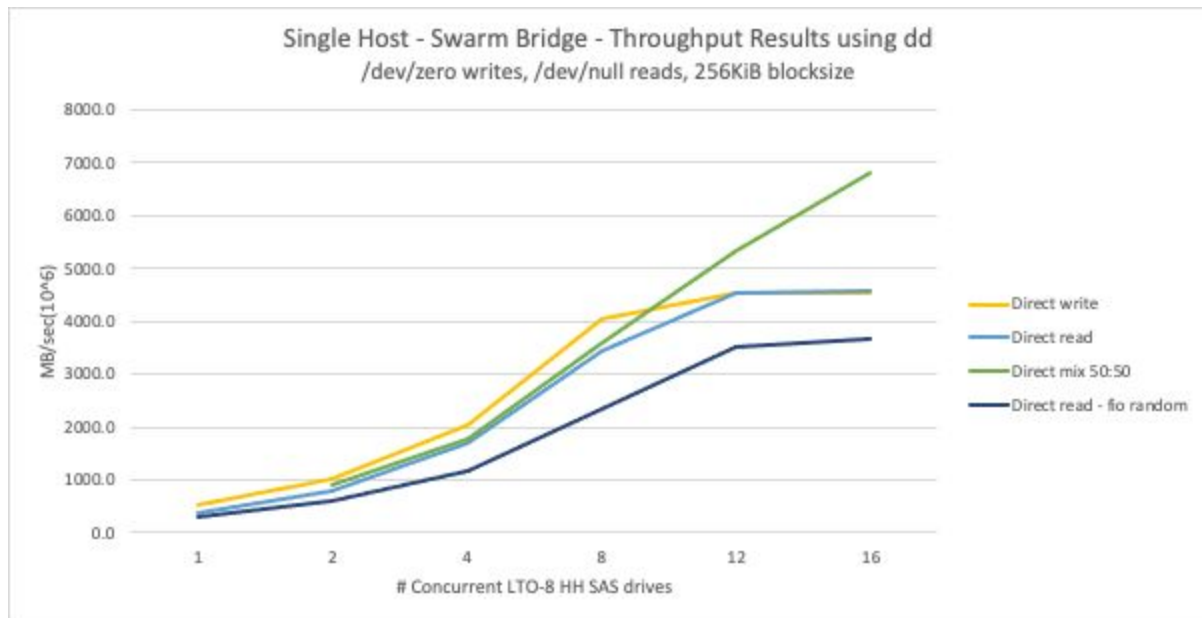
Reads are unusual in that the drive average increases as you scale drives. 1 drive reads at 370.5 MB/sec, but with 8 drives, they all read at 422.3. Repeatable.

Data from same run, but with scaled performance.

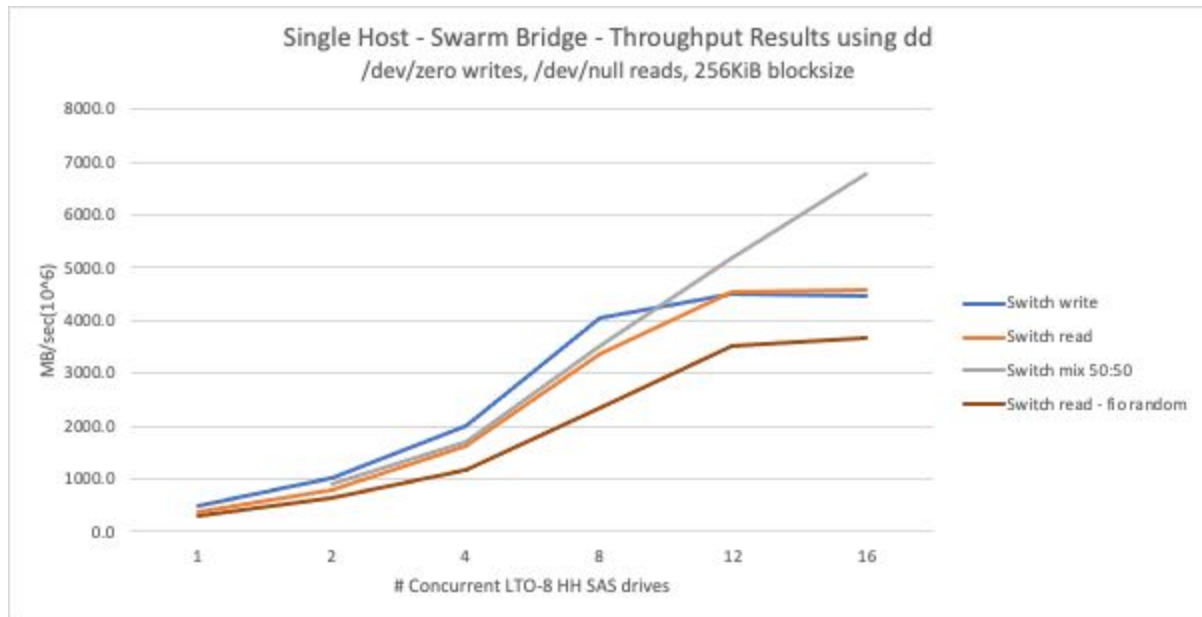


Once again reads are experiencing differences in performance behaviour and should be investigated.

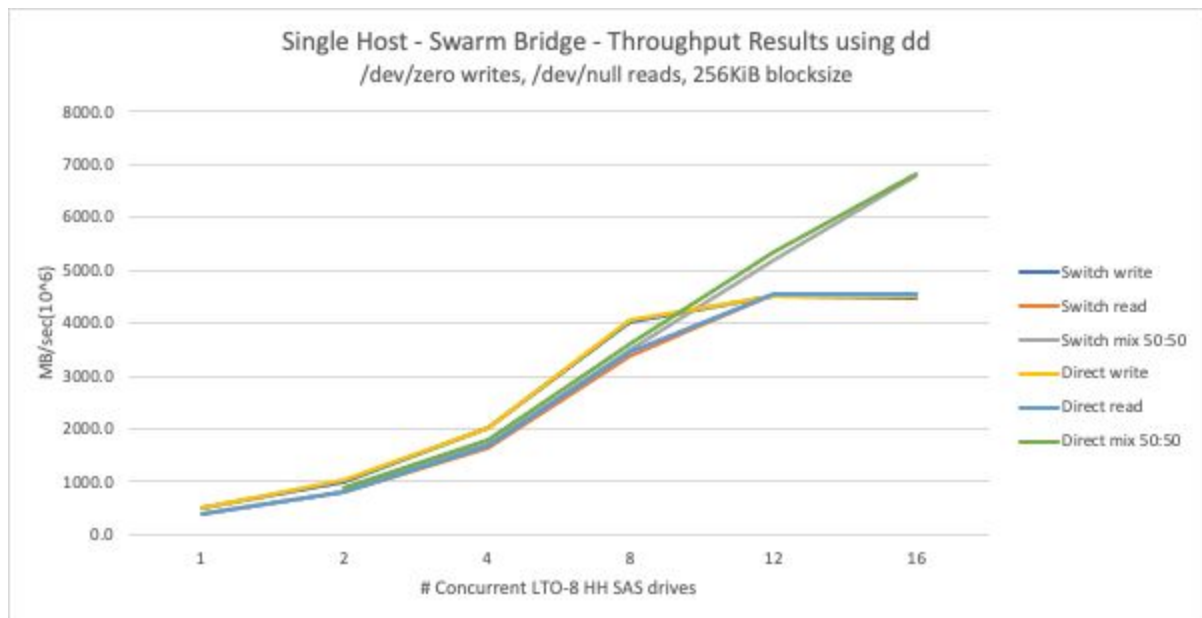
Direct connect results



Switch connect results

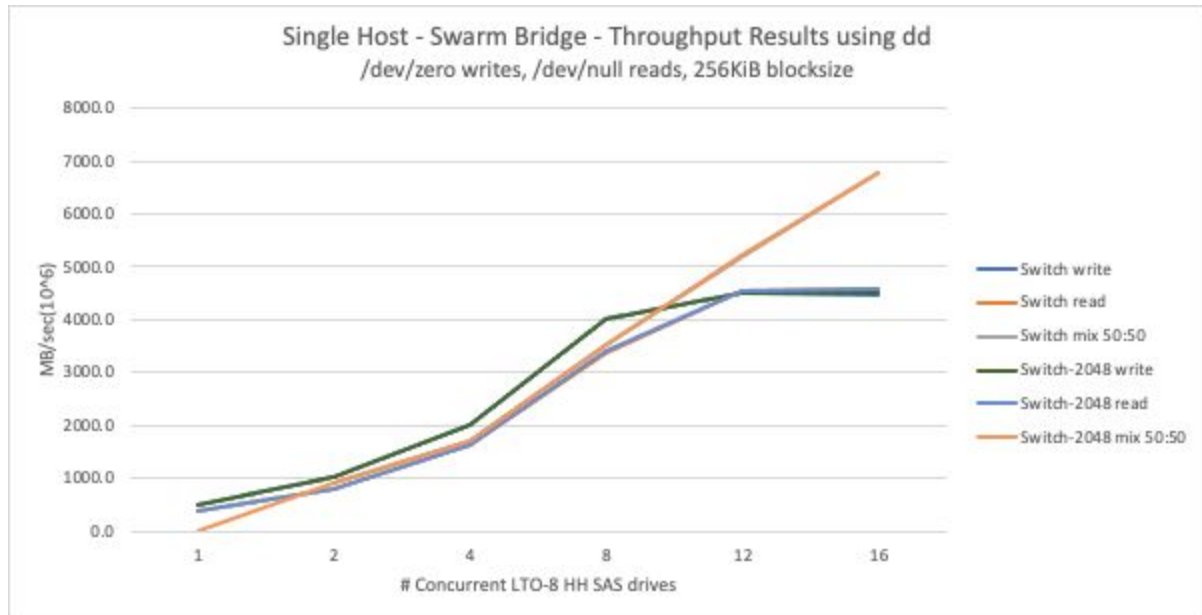


Switch vs Direct



No significant difference measured when comparing direct connect versus switch connect.

Max_sectors 1024 vs 2048



No significant difference between `ib_iser` module `max_sectors` at default of 1024 vs setting to 2048.

Check with `'modprobe -c | grep ib_iser'`
`options ib_iser max_sectors=2048`

Spectrum Protect

IBM Spectrum Protect server instance was installed on the host. Swarm hosted tape devices were picked up without problem by the IBM device driver. The tape devices were able to be configured to the server instance. Tapes were checked in and labeled. A storage pool and device class were created and a client backup successfully wrote to the tape storage.

Truncated device configuration:

/* Device Configuration */

```
DEFINE DEVCLASS LTO_CLASS DEVT=LTO FORMAT=ULTRIUM8C MOUNTL=DRIVES MOUNTWAIT=60
MOUNTRETENTION=60 PREFIX=ADSM LIBRARY=SWARMLIB WORM=NO DRIVEENCRYPTION=ALLOW
LBPROTECT=NO
SET SERVERNAME SERVER1
DEFINE LIBRARY SWARMLIB LIBTYPE=SCSI SERIAL="926000XXXX" SHARED=NO AUTOLABEL=NO
RESETDRIVE=NO
DEFINE DRIVE SWARMLIB LTO0 ELEMENT=261 ONLINE=Yes WWN="21620090A5003287"
SERIAL="106200XXXX"
DEFINE DRIVE SWARMLIB LTO1 ELEMENT=260 ONLINE=Yes WWN="21610090A5003287"
SERIAL="106100XXXX"
.....

/* LIBRARYINVENTORY SCSI SWARMLIB DIAG01L8 4096 101*/
/* LIBRARYINVENTORY SCSI SWARMLIB DIAG02L8 4097 101*/
..... . .

DEFINE PATH SERVER1 SWARMLIB SRCTYPE=SERVER DESTTYPE=LIBRARY DEVICE=/dev/tmscsi/lb0
ONLINE=YES
DEFINE PATH SERVER1 LTO0 SRCTYPE=SERVER DESTTYPE=DRIVE LIBRARY=SWARMLIB
DEVICE=/dev/IBMtape0 ONLINE=YES
DEFINE PATH SERVER1 LTO1 SRCTYPE=SERVER DESTTYPE=DRIVE LIBRARY=SWARMLIB
DEVICE=/dev/IBMtape1 ONLINE=YES
..... .

SERVERBACKUPNODEID 1
```

Summary

Overall the swarm appliance POC was successful in demonstrating a viable replacement for fibre channel.

Performance

Compressed streaming performance would have higher achievable performance with traditional fibre channel or SAS connections. LTO-8 Half-height drives have a lower native data rate than full height drives. With LTO9 there will be a full height SAS option that would be more compelling in terms of performance and reliability. Tape motion operations should be identical: mounts, loads, unloads, seeks, etc... Unlikely to see a performance difference in normal operations.

Configuration

Configuration steps on the Swarm appliance, switch, tape library, and host were different enough from fibre channel to require assistance to configure for the first time. Once the commands and differences were known, it was simple to refresh device configuration on the host.

Troubleshooting

Encountered a performance problem where pause frames were not being honored. Arista switch had priority flow control(PFC) configured per the Mellanox web site. Took a number of weeks to find a work around by disabling PFC on the switch and setting up global priority flow control.

Redundancy

Require two Swarm appliances to provide redundancy. If a major issue impacted the swarm appliance, you would lose access to up to 16 tape drives. Unlikely to be able to bypass the appliance and directly connect to the tape drive since you would require a SAS adapter in the host and be within distance to the tape drive for the SAS cable.

Cost benefit analysis

Without having a side by side comparison of all the components, unable to declare a winner in this area. Swarm appliance does reduce the number of optical connections significantly and you could leverage existing ethernet switch infrastructure. Since you can connect 16 SAS tape drives to the appliance, you would reduce the number of optical connections from the tape drives from 16 to the 2 ethernet connections out of the swarm appliance. Host HBA likely a wash in price from a fibre channel HBA to a RDMA RoCE capable converged ethernet HBA.